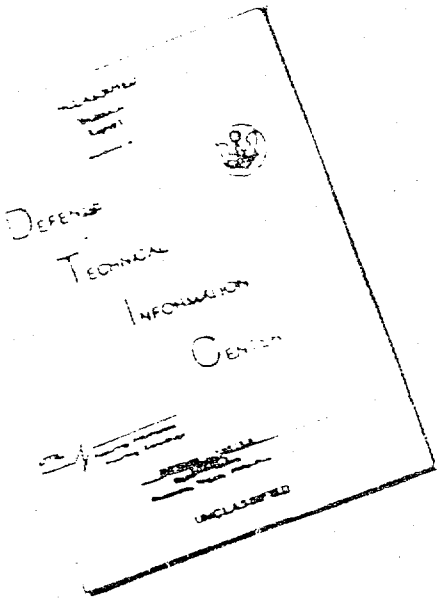
AD 742102

Best Available Copy

DISCLAIMER NOTICE



THIS DOCUMENT IS BEST QUALITY AVAILABLE. THE COPY FURNISHED TO DTIC CONTAINED A SIGNIFICANT NUMBER OF PAGES WHICH DO NOT REPRODUCE LEGIBLY.

CONTROL ANALYSIS CORPORATION
900 Welch Road
Palo Alto, California

Technical Report No. 86-1
April 21, 1972

A NEW APPROACH TO SIMULATING STABLE STOCHASTIC SYSTEMS: I - GENERAL MULTI-SERVER QUEUES

by
Michael A. Crane and Donald L. Iglehart

This research was sponsored by the Office of Naval Research under contract N00014-72-C-0086 [NR-047-106].

Unclassified

Security Classification

DOCUMENT CONTROL DATA - R&D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1. ORIGINATING ACTIVITY (Corporate author) Control Analysis Corporation 900 Welch Road Palo Alto, California	2a. REPORT SECURITY CLASSIFICATION Unclassified 2b. GROUP

3. REPORT TITLE

A New Approach to Simulating Stable Stochastic Systems: I - General Multi-Server Queues

4. DESCRIPTIVE NOTES (Type of report and inclusive dates)

Technical Report

5. AUTHOR(S) (Last name, first name, initial)

CRANE, Michael A., and IGLEHART, Donald L.

6. REPORT DATE	7a. TOTAL NO. OF PAGES	7b. NO. OF REFS
April 21, 1972	41	9

8a. CONTRACT OR GRANT NO.	9a. ORIGINATOR'S REPORT NUMBER(S)
N00014-72-C-0086 b. PROJECT NO. NR-047-106 c. d.	Technical Report No. 86-1 9b. OTHER REPORT NO(S) (Any other numbers that may be assigned this report)

10. AVAILABILITY/LIMITATION NOTICES

Distribution of this document is unlimited

11. SUPPLEMENTARY NOTES	12. SPONSORING MILITARY ACTIVITY
	Operations Research Program Office of Naval Research Arlington, Virginia 22217

13. ABSTRACT

A new technique is introduced for analyzing simulations of stochastic systems in the steady state. From the viewpoint of classical statistics, questions of simulation run duration and of starting and stopping simulations are addressed. This is possible because of the existence of a random grouping of observations which produces independent identically distributed blocks from the start of the simulation.

The analysis is presented in the context of the general multi-server queue, with arbitrarily distributed inter-arrival and service times. In this case, it is the busy period structure of the system which produces the grouping mentioned above. Numerical illustrations are given for the M/M/1 queue. Statistical methods are employed so as to obtain confidence intervals for a variety of parameters of interest, such as the expected value of the stationary customer waiting time, the expected value of a function of the stationary waiting time, the expected number of customers served and length of a busy cycle, the tail of the stationary waiting time distribution, and the standard deviation of the stationary waiting time. Consideration is also given to determining system sensitivity to errors and uncertainty in the input parameters.

DD FORM 1 JAN 64 1473

Unclassified

Security Classification

UNCLASSIFIED

Security Classification

14. KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
simulation						
queueing simulation						
statistical analysis of simulations						
confidence intervals in simulation						
estimation						
sampling						

INSTRUCTIONS

1. **ORIGINATING ACTIVITY:** Enter the name and address of the contractor, subcontractor, grantee, Department of Defense activity or other organization *(corporate author)* issuing the report.

2a. **REPORT SECURITY CLASSIFICATION:** Enter the overall security classification of the report. Indicate whether "Restricted Data" is included. Marking is to be in accordance with appropriate security regulations.

2b. **GROUP:** Automatic downgrading is specified in DoD Directive 5200.10 and Armed Forces Industrial Manual. Enter the group number. Also, when applicable, show that optional markings have been used for Group 3 and Group 4 as authorized.

3. **REPORT TITLE:** Enter the complete report title in all capital letters. Titles in all cases should be unclassified. If a meaningful title cannot be selected without classification, show title classification in all capitals in parenthesis immediately following the title.

4. **DESCRIPTIVE NOTES:** If appropriate, enter the type of report, e.g., interim, progress, summary, annual, or final. Give the inclusive dates when a specific reporting period is covered.

5. **AUTHOR(S):** Enter the name(s) of author(s) as shown on or in the report. Enter last name, first name, middle initial. If military, show rank and branch of service. The name of the principal author is an absolute minimum requirement.

6. **REPORT DATE:** Enter the date of the report as day, month, year; or month, year. If more than one date appears on the report, use date of publication.

7a. **TOTAL NUMBER OF PAGES:** The total page count should follow normal pagination procedures, i.e., enter the number of pages containing information.

7b. **NUMBER OF REFERENCES:** Enter the total number of references cited in the report.

8a. **CONTRACT OR GRANT NUMBER:** If appropriate, enter the applicable number of the contract or grant under which the report was written.

8b, 8c, & 8d. **PROJECT NUMBER:** Enter the appropriate military department identification, such as project number, subproject number, system numbers, task number, etc.

9a. **ORIGINATOR'S REPORT NUMBER(S):** Enter the official report number by which the document will be identified and controlled by the originating activity. This number must be unique to this report.

9b. **OTHER REPORT NUMBER(S):** If the report has been assigned any other report numbers *(either by the originator or by the sponsor)*, also enter this number(s).

10. **AVAILABILITY/LIMITATION NOTICES:** Enter any limitations on further dissemination of the report, other than those imposed by security classification, using standard statements such as:

(1) "Qualified requesters may obtain copies of this report from DDC."

(2) "Foreign announcement and dissemination of this report by DDC is not authorized."

(3) "U. S. Government agencies may obtain copies of this report directly from DDC. Other qualified DDC users shall request through _____."

(4) "U. S. military agencies may obtain copies of this report directly from DDC. Other qualified users shall request through _____."

(5) "All distribution of this report is controlled. Qualified DDC users shall request through _____."

If the report has been furnished to the Office of Technical Services, Department of Commerce, for sale to the public, indicate this fact and enter the price, if known.

11. **SUPPLEMENTARY NOTES:** Use for additional explanatory notes.

12. **SPONSORING MILITARY ACTIVITY:** Enter the name of the departmental project office or laboratory sponsoring *(paying for)* the research and development. Include address.

13. **ABSTRACT:** Enter an abstract giving a brief and factual summary of the document indicative of the report, even though it may also appear elsewhere in the body of the technical report. If additional space is required, a continuation sheet shall be attached.

It is highly desirable that the abstract of classified reports be unclassified. Each paragraph of the abstract shall end with an indication of the military security classification of the information in the paragraph, represented as *(TS), (S), (C), or (U)*.

There is no limitation on the length of the abstract. However, the suggested length is from 150 to 225 words.

14. **KEY WORDS:** Key words are technically meaningful terms or short phrases that characterize a report and may be used as index entries for cataloging the report. Key words must be selected so that no security classification is required. Identifiers, such as equipment model designation, trade name, military project code name, geographic location, may be used as key words but will be followed by an indication of technical context. The assignment of links, roles, and weights is optional.

TABLE OF CONTENTS

NONTECHNICAL SUMMARY

In this paper, we introduce a new technique for analyzing simulations of stochastic systems in the steady state. From the viewpoint of classical statistics, we address the questions of simulation run duration and of starting and stopping simulations. We are able to do so by avoiding two difficulties which have previously made classical statistics inappropriate for simulation analyses. These are the statistical dependence between successive observations and the inability of the simulator to begin the system in the steady state.

For many stochastic systems being simulated it is possible to find a random grouping of observations which produces independent identically distributed blocks from the start of the simulation. This grouping then enables the simulator to avoid the two problems mentioned above. He has at his disposal the methods of classical statistics such as confidence intervals, hypothesis testing, regression, and sequential estimation which are appropriate for independent observations. Furthermore, information that is useful in estimating the steady-state behavior of the system can be collected from scratch thus eliminating the problem of the initial transient.

The approach mentioned above is appropriate for the simulation of systems which returns infinitely often to a single state. In the current paper, we restrict our discussion to the general multi-server queue, with arbitrarily distributed inter-arrival and service times. We leave the more general systems to future publications.

Section 2 reviews some relevant results from the theory of statistical inference. Sections 3 and 4 discuss the probabilistic structure of stable queueing systems. A queueing system is stable if the customer arrival rate is strictly less than the maximum service rate (with all servers working). In this case, it may be shown under mild conditions that the idle state (the state in which all servers are idle) occurs infinitely often. Furthermore, letting a busy cycle refer to the time interval between two successive idle states, it may be shown that observations made in different busy cycles are statistically independent and identically distributed. These observations might be, for example, the number of customers served in the busy cycle, the length of the busy cycle, the sum of the waiting times of customers served in the busy cycle, or the sum of some function of the waiting times of the customers served in the busy cycle.

It is shown that each busy cycle provides information relating to the system's steady-state behavior. In particular, the expected value of any well-behaved function of the steady-state waiting time is equal to the expected value of the sum of that same function of the individual customer waiting times in a single busy cycle, divided by the expected number of customers served in a busy cycle. This fact may be used with the above-mentioned independence to enable the simulator to perform a thorough statistical analysis of the steady state. One merely directs his analysis toward the estimation of properties of the individual busy cycle and then infers corresponding properties for the steady state. The former task is simplified because of the independence and identical probabilistic structure of different busy cycles, which permits classical statistical analysis.

To illustrate the application of these ideas, consider the problem of obtaining a confidence interval for the expected steady-state waiting time, $E\{W\}$. Let Y_k denote the sum of the customer waiting times in the kth busy cycle and let $\bar{Y}$ and s_Y^2 denote respectively the sample mean and sample variance for Y_k in N observations (busy cycles). Let α_k denote the number of customers served in the kth busy cycle, and let $\bar{\alpha}$ and s_α^2 denote respectively the sample mean and sample variance for α_k in N observations. To obtain a confidence interval for $E\{W\}$ with at least $100(1-\gamma)\%$ confidence, we observe the system for N busy cycles. The interval computed is then

$$\frac{\max\{0,\ \bar{Y}-z_{1-\gamma_1/2}\ s_Y/\sqrt{N}\}}{\bar{\alpha}+z_{1-\gamma_2/2}\ s_\alpha/\sqrt{N}} \leq E\{W\} \leq \frac{\bar{Y}+z_{1-\gamma_1/2}\ s_Y/\sqrt{N}}{\max\{0,\ \bar{\alpha}-z_{1-\gamma_2/2}\ s_\alpha/\sqrt{N}\}}$$

where z_x is the $100x$ percentile for the normal distribution and γ_1 and γ_2 satisfy $\gamma_1+\gamma_2=\gamma$.

In Section 5, a two-step procedure is developed for obtaining an approximate confidence interval for the steady-state mean of a general function of the customer waiting time. The first step consists of a short simulation run which serves as a planning guide for a much longer second run. At the end of the first run, the simulator is able to estimate the minimum confidence interval length which can be obtained from the second run given a specified level of confidence and run time. The procedure is thus viewed as a tool by which the simulator may balance computation cost against level

of precision and hence implement a rational design of experiment.

In Section 6, we illustrate the application of the statistical techniques to an actual queueing simulation. For illustration purposes, we choose the single-server queue with exponential inter-arrival and service time distributions, since its theoretical properties are well-known and provide a means of comparison with simulation results. Confidence intervals are obtained for various steady-state quantities of interest, including the mean and standard deviation of the waiting time, the expected number of customers in the system, the expected value of a penalty function on the waiting time, the tail of the waiting time distribution, the expected number of customers served in a busy cycle, the expected value of a penalty function on the number of customers served in a busy cycle, the tail of the distribution for the number of customers served in a busy cycle, and others. The two-step procedure is then illustrated for the mean waiting time. In a single realization of a second-step run consisting of 18,000 busy cycles, a 98% confidence interval [.086, .108] is obtained for a queue with an arrival rate of 5 customers per time unit and a service rate of 10 customers per time unit (the theoretical mean waiting time is 0.1). Two systems with different service rates are then compared by means of a confidence interval on the difference of the mean waiting times. Finally, a method is illustrated by which one may infer sensitivity to an unknown parameter. In particular, if the arrival rate is unknown and one observes the system at two different values for the arrival rate, then under certain linearizing assumptions, one may make

confidence interval statements for any arrival rate between the two observed values. Indeed, one may obtain a confidence band about the expected waiting time (or some function of the waiting time) expressed as a function of the unknown parameter. It is felt that such a capability could be an effective aid in assessing the adequacy of the input data accuracy, and in efficiently uncovering those parameters that warrant more in-depth analysis. Of course this technique, together with all of the techniques illustrated in Section 6, is only feasible because of the busy period structure which is exploited.

Finally, in Section 7, we compare our approach for estimating the mean steady-state waiting time with a commonly used alternative. In our approach, one arrives at an estimate essentially by forming the sample mean of the customer waiting times in some fixed number of busy cycles. The actual number of customers observed is thus random. In an alternative approach, one might simulate a fixed number of customers, disregard an initial fraction of the customers as transient observations, then form a sample mean with the remaining customers. In both cases, the estimate obtained converges to the theoretical value as the number of observations becomes large. In our approach, however, one is able to make rigorous statistical statements by taking advantage of the independence of the busy cycles. In Section 7, we seek to determine whether this advantage is obtained at the expense of reduced accuracy in the estimator. Experimental results are obtained for the single-server queue used in the previous illustrations. It is found that for this queue, statistics which are

obtained in comparable length computer runs for these two approaches are very close in accuracy. Thus, one may apparently make use of the methods of this report without fear of reduced accuracy.

Continuing research in this area will extend and illustrate this approach for more general simulations and describe in detail the steps necessary for implementation.

A NEW APPROACH TO SIMULATING STABLE STOCHASTIC SYSTEMS: I - GENERAL MULTI-SERVER QUEUES

by

Michael A. Crane and Donald L. Iglehart

1. INTRODUCTION

The principal goal of most simulations of stable stochastic systems is to estimate properties of the stationary or steady-state behavior of the system. Two of the major problems in such simulations are the statistical dependence between successive observations and the inability of the simulator to begin the system in the steady-state. The first problem has necessitated using methods of time series analysis rather than classical statistics. The second has inspired many simulators to let the system run for a sufficient length of time so that the initial transient wears off and a steady-state condition obtains. This procedure, of course, requires a judgement on how long to let the system run before making observations.

For many stochastic systems being simulated it is possible to find a random grouping of observations which produces independent identically distributed (i.i.d.) blocks from the start of the simulation. This grouping then enables the simulator to avoid the two problems mentioned above. He has at his disposal the methods of classical statistical analysis such as confidence intervals, hypothesis testing, regression, and sequential estimation since the

observations are now i.i.d. Furthermore, information that is useful in estimating the steady-state behavior of the system can be collected from scratch thus eliminating the problem of the initial transient.

The key requirement for obtaining these i.i.d. blocks is that the system being simulated return to a single state infinitely often and that the mean time between such returns is finite. This requirement will be met for many, but not all, stable systems that might be simulated.

In this paper we shall illustrate the main ideas of this approach in the context of the GI/G/s queue, where $s \geq 1$. Other stochastic systems which can be dealt with in this same manner will be discussed in future publications. This paper is organized as follows. Section 2 reviews some revelant results from the theory of statistical inference. Section 3 summarizes the probabilistic structure of the GI/G/1 queue with an eye toward using these results in carrying out a simulation. In Section 4 a similar treatment is given to the GI/G/s queue for $s > 1$. In Section 5 we discuss a two-step procedure for estimating the stationary expected value of a general function of the waiting time. Numerical illustrations for the M/M/1 queue are given in Section 6. Finally, in Section 7 we compare our procedure with a commonly used alternative.

2. PRELIMINARY RESULTS

Before proceeding further, it is useful to review some relevant ideas from the theory of statistical inference. As usual, we have a given probability triple and a sequence of random variables (r.v.'s)

defined on it. Since the detailed construction of this triple is not important, we omit it from the discussion. Let θ be some unknown parameter and let $\underline{\theta}$ and $\bar{\theta}$ be two random variables obtained from a statistical experiment. If $0 < \gamma < 1$ and $P\{\underline{\theta} \leq \theta \leq \bar{\theta}\} \geq 1 - \gamma$, we say that $[\underline{\theta}, \bar{\theta}]$ is a confidence interval for θ with at least $100(1-\gamma)\%$ confidence. Roughly speaking, if many values are obtained for $\underline{\theta}$ and $\bar{\theta}$ in independent replications of the experiment, the interval $[\underline{\theta}, \bar{\theta}]$ will surround θ at least $100(1-\gamma)\%$ of the time.

Given confidence intervals for one or more parameters, it is possible to obtain confidence intervals for various functions of these parameters. For example, suppose θ_1 and θ_2 are two unknown parameters, and $[\underline{\theta}_1, \bar{\theta}_1]$ and $[\underline{\theta}_2, \bar{\theta}_2]$ are confidence intervals for θ_1 and θ_2 with at least $100(1-\gamma_1)\%$ and $100(1-\gamma_2)\%$ confidence, that is $P\{\underline{\theta}_1 \leq \theta_1 \leq \bar{\theta}_1\} \geq 1-\gamma_1$, and $P\{\underline{\theta}_2 \leq \theta_2 \leq \bar{\theta}_2\} \geq 1-\gamma_2$. Then by a straightforward argument,

$$P\{\underline{\theta}_1 - \bar{\theta}_2 \leq \theta_1 - \theta_2 \leq \bar{\theta}_1 - \underline{\theta}_2\} \geq 1-\gamma_1-\gamma_2. \tag{1}$$

If $A > 0$, then

$$P\{A\underline{\theta}_1 + B \leq A\theta_1 + B \leq A\bar{\theta}_1 + B\} \geq 1 - \gamma_1. \tag{2}$$

If $\theta_1 \geq 0$ and $a > 0$, then

$$P\{([\underline{\theta}_1]^+)^a \leq \theta_1^a \leq ([\bar{\theta}_1]^+)^a\} \geq 1 - \gamma_1, \tag{3}$$

where $[x]^+ = \max(0,x)$. If $\theta_1 \geq 0$ and $\theta_2 > 0$, then

$$P\{\frac{[\underline{\theta}_1]^+}{\bar{\theta}_2} \leq \frac{\theta_1}{\theta_2} \leq \frac{\bar{\theta}_1}{[\underline{\theta}_2]^+}\} \geq 1 - \gamma_1 - \gamma_2. \tag{4}$$

Finally, suppose $\theta(\lambda)$ is an unknown parameter which is a linear function of the parameter λ. Suppose $\lambda_1 < \lambda_2$ and $P\{\underline{\theta} \leq \theta(\lambda_1) \leq \bar{\theta}_1\} \geq 1 - \gamma_1$ and $P\{\underline{\theta}_2 \leq \theta(\lambda_2) \leq \bar{\theta}_2\} \geq 1 - \gamma_2$. Then

$$P\{\underline{\theta}_1 + \frac{(\lambda-\lambda_1)(\underline{\theta}_2 - \underline{\theta}_1)}{\lambda_2 - \lambda_1} \leq \theta(\lambda) \leq \bar{\theta}_1 + \frac{(\lambda-\lambda_1)(\bar{\theta}_2 - \bar{\theta}_1)}{\lambda_2 - \lambda_1}$$

$$\text{for all } \lambda_1 \leq \lambda \leq \lambda_2\} \geq 1 - \gamma_1 - \gamma_2. \tag{5}$$

This fact is potentially useful for studying the sensitivity of a <u>nonlinear</u> function $\theta(\lambda)$ to λ over some small interval $[\lambda_1, \lambda_2]$.

Now let $\{X_n : n \geq 1\}$ be a sequence of i.i.d. r.v.'s with finite second moments, and define $\theta = E\{X_1\}$ and $\sigma^2 = \sigma^2\{X_1\}$. Let Φ denote the distribution function for a normal random variable with zero mean and variance one; i.e.,

$$\Phi(x) = \int_{-\infty}^{x} \phi(\xi)d\xi \qquad -\infty < x < \infty,$$

where

$$\phi(\xi) = \frac{1}{\sqrt{2\pi}} e^{-\xi^2/2} \qquad -\infty < \xi < \infty.$$

For $0 < x < 1$, define $z_x \equiv \Phi^{-1}(x)$. Also, define the sample mean

and sample standard deviation

$$\bar{X}(n) \equiv \frac{1}{n} \sum_{j=1}^{n} X_j \qquad n=1,2,\ldots,$$

$$s(n) \equiv [\frac{1}{n-1} \sum_{j=1}^{n} (X_j - \bar{X}(n))^2]^{1/2}, \quad n=2,3,\ldots.$$

Now we know by the central limit theorem that the

$$\lim_{n\to\infty} P\{\sqrt{n}\ (\bar{X}(n) - \theta)/\sigma \leq z\} = \Phi(z), \ -\infty < z < \infty.$$

Furthermore, it may be shown using the strong law of large numbers that the

$$P\{\lim_{n\to\infty} s(n) = \sigma\} = 1.$$

It then follows by a well-known result, cf. CHUNG (1968), Theorem 4.4.8, that the

$$\lim_{n\to\infty} P\{\sqrt{n}(\bar{X}(n) - \theta)/s(n) \leq z\} = \Phi(z), \ -\infty < z < \infty$$

Hence, for $0 < \gamma < 1$ and large n,

$$P\{-z_{1-\gamma/2} \leq \sqrt{n}(\bar{X}(n) - \theta)/s(n) \leq z_{1-\gamma/2}\} \approx 1-\gamma$$

or

$$P\{\bar{X}(n) - z_{1-\gamma/2}s(n)/\sqrt{n} \leq \theta \leq \bar{X}(n) + z_{1-\gamma/2}s(n)/\sqrt{n}\} \approx 1-\gamma,$$

giving an approximate $100(1-\gamma)\%$ confidence interval for θ. This method will be used for analyzing sequences of i.i.d. r.v.'s introduced in later sections.

3. PROBABILISTIC STRUCTURE OF THE GI/G/1 QUEUE

Consider now a GI/G/1 queueing system in which the 0th customer arrives at time $t_0 = 0$, finds a free server, and experiences a service time v_0. The nth customer arrives at time t_n and experiences a service time v_n. Let the interarrival times $t_n - t_{n-1} = u_n$, $n \geq 1$. Assume that the two sequences $\{v_n : n \geq 0\}$ and $\{u_n : n \geq 1\}$ each consist of i.i.d. r.v.'s and are themselves independent. Let $E\{u_n\} = \lambda^{-1}$, $E\{v_n\} = \mu^{-1}$, and $\rho = \lambda/\mu$ where $0 < \lambda, \mu < \infty$. Thus μ (λ) has the interpretation of the mean service (arrival) rate. The parameter ρ is called the traffic intensity and is the natural measure of congestion for this system. We shall assume that $\rho < 1$, a necessary and sufficient condition for the system to be stable.

The principal system characteristics of interest are $Q(t)$, the number of customers in the system at time t; W_n, the waiting time (time from arrival to commencement of service) of the nth customer; $W(t)$, the work load facing the server at time t; $B(t)$, the amount of time in the interval $[0,t]$ that the server is busy; and $D(t)$, the total number of customers who have been served and have departed from the system in $[0,t]$.

Here we shall review the basic structure of the GI/G/1 queue relevant to our simulation study. For a comprehensive treatment of

these and other results for the GI/G/1 queue see IGLEHART (1971). To begin the analysis of the process $\{W_n : n \geq 0\}$ let $X_n = v_{n-1} - u_n$ and set $S_0 = 0$, $S_n = X_1 + \ldots + X_n$, $n \geq 1$. The following recursive relationship exists for the W_n's:

$$W_0 = 0$$

$$W_{n+1} = [W_n + X_{n+1}]^+, \quad n \geq 0.$$

By induction one can show that

$$W_n = \max\{S_n - S_k : k = 0,1,\ldots,n\}, \quad n \geq 0.$$

Using the notion of optional r.v.'s, it can be shown that there exists a sequence of r.v.'s $\{\beta_k : k \geq 0\}$ such that $\beta_0 = 0$, $\beta_k < \beta_{k+1}$, and $W_{\beta_k} = 0$ with probability one. In other words, the customers numbered β_k are those lucky fellows who arrive to find a free server and experience of no waiting in the queue. The fact that there exists an infinite number of such customers is a direct consequence of the assumption that $\rho < 1$. The time axis $R_+^1 = [0,\infty)$ can be divided into alternating intervals during which the server is busy, idle, busy, etc. We call these intervals busy periods (b.p.'s) and idle periods (i.p.'s). An i.p. plus the preceding b.p. is called a busy cycle (b.c.). If we let $\alpha_k = \beta_k - \beta_{k-1}$, $k \geq 1$, then α_k represents the number of customers served in the kth busy period (b.p.) and they are numbered $\{\beta_{k-1}, \beta_{k-1} + 1, \ldots, \beta_k - 1\}$.

Next define the random vectors $\underset{\sim}{X}_k = (v_{k-1}, u_k)$ and $\underset{\sim}{Y}_k = \{\alpha_k, \underset{\sim}{X}_{\beta_{k-1}+1}, \ldots, \underset{\sim}{X}_{\beta_k}\}$, $k \geq 1$. Note that these vectors are unusual in the sense that they have a random number of components, namely, α_k+1. Observe that the vector $\underset{\sim}{Y}_1 = \{\alpha_1, \underset{\sim}{X}_1, \ldots, \underset{\sim}{X}_{\alpha_1}\}$ includes all the data required to completely construct the behavior of the system in the first b.p. The principal fact that permits us to decompose the system into i.i.d. blocks is that the $\underset{\sim}{Y}_k$'s are i.i.d. Hence we have the intuitively plausible conclusion that comparable r.v.'s in different b.p.'s are i.i.d. Now define the following r.v.'s for $k \geq 1$:

$$\eta_k = v_{\beta_{k-1}} + \ldots + v_{\beta_k - 1},$$

$$\xi_k = u_{\beta_{k-1}+1} + \ldots + u_{\beta_k},$$

$$\nu_k = \xi_k - \eta_k,$$

$$Y_k^{(1)} = \sum_{j=0}^{\alpha_k - 1} (S_{\beta_{k-1}+j} - S_{\beta_{k-1}}),$$

$$Y_k^{(2)} = Y_k^{(1)} + \eta_k, \qquad \text{and}$$

$$Y_k^{(3)} = \sum_{j=0}^{\alpha_k - 1} \{(S_{\beta_{k-1}+j} - S_{\beta_{k-1}}) v_{\beta_{k-1}+j} + \frac{1}{2} v^2_{\beta_{k-1}+j}\}.$$

These r.v.'s have the following interpretations: η_k is the length of the kth b.p., ξ_k the length of the kth b.c., ν_k the length of the kth i.p., $Y_k^{(1)}$ the sum of the waiting times in

kth b.p., $Y_k^{(2)}$ the integral under the curve $Q(s)$ in the kth b.p., and $Y_k^{(3)}$ the integral under the curve $W(s)$ in the kth b.p. Because the $\underset{\sim}{Y}_k$'s are i.i.d. so are the η_k's, ξ_k's, ν_k's, $Y_k^{(1)}$'s, $Y_k^{(2)}$'s, and $Y_k^{(3)}$'s. It is these i.i.d. sequences that we shall observe in our simulation with an eye toward using classical statistical methods of analysis.

Next we record some known results on expected values of these r.v.'s. We shall assume that the $E\{v_0^2\} < \infty$. Then the $E\{\alpha_k\} = \exp\{\sum_{n=1}^{\infty} n^{-1} P[S_n > 0]\}$, $E\{\eta_k\} = \mu^{-1} E\{\alpha_k\}$, $E\{\xi_k\} = \lambda^{-1} E\{\alpha_k\}$, $E\{\nu_k\} = \lambda^{-1}(1-\rho) E\{\alpha_k\}$, $E\{Y_k^{(1)}\} = E\{W\} E\{\alpha_k\} < \infty$, $E\{Y_k^{(2)}\} = (\lambda E\{W\} + \rho) E\{\xi_k\}$, and $E\{Y_k^{(3)}\} = (\rho E\{W\} + \frac{1}{2}\lambda E\{v_0^2\}) E\{\xi_k\}$, where W is the r.v. to which W_n converges in distribution. In other words, W is the so-called stationary waiting time. An expression for the characteristic function of W exists, in terms of the $E\{e^{itS_n^+}\}$, however, it is difficult to evaluate for specific cases. Also $E\{W\} = \sum_{n=1}^{\infty} n^{-1} E\{S_n^+\}$. If u_1 has a non-lattice distribution, it is known that $Q(t) \Rightarrow Q$ and $W(t) \Rightarrow W^*$ as $t \to \infty$, where $\Rightarrow$ denotes weak convergence (convergence in distribution). Hence Q is the stationary queue length and W^* is the stationary virtual waiting time. Furthermore, $E\{Q\} = \lambda E\{W\} + \rho$ and $E\{W^*\} = \rho E\{W\} + \lambda E\{v_0^2\}/2$ which are the terms appearing in the expressions for $E\{Y_k^{(2)}\}$ and $E\{Y_k^{(3)}\}$.

More generally let f be a measurable function from $[0,\infty)$ into $(-\infty,\infty)$. Assuming that the $E\{f(W)\} < \infty$, then $E\{\sum_{j=0}^{\alpha_k - 1} f(S_{\beta_{k-1}+j} - S_{\beta_{k-1}})\} = E\{f(W)\} E\{\alpha_k\}$. This identity plus the fact that the $\underset{\sim}{Y}_k$'s are i.i.d. allows us to treat $E\{f(W)\}$ in the

same manner that we do $E\{W\}$. In Section 6 a variety of functions f of practical interest will be discussed.

While the parameters $E\{\alpha_1\}$, $E\{W\}$, $E\{Q\}$, $E\{W^*\}$, and $E\{f(W)\}$ can theoretically all be calculated from the distributions of v_0 and u_1, these calculations are very difficult and one might choose to estimate them through simulation. It is problems of this sort that we address in this paper.

4. PROBABILISTIC STRUCTURE OF THE GI/G/s QUEUE

Now suppose there are $s > 1$ servers. The appropriate value of the traffic intensity is now $\rho = \lambda/s\mu$, which we assume again is strictly less than one. It is known (see KIEFER and WOLFOWITZ (1955, 1956), and LOYNES (1962)) that $\rho < 1$ is a necessary and sufficient condition for the queueing system to be stable. However, $\rho < 1$ is not enough to insure that the queue becomes idle infinitely often, which is the key to the analysis in the case $s = 1$. A simple and reasonable condition, which we shall henceforth assume, was recently provided by WHITT (1971); namely, that the $P\{u_n > v_{n-1}\} > 0$. Additional results on this problem are contained in KENNEDY (1971).

The analysis of this queue is best approached through a vector-valued waiting time process $\{\underset{\sim}{W}_n : n \geq 0\}$ introduced by KIEFER and WOLFOWITZ (1955, 1956). We assume the customers are served by the first available server. If more than one server is free, the customer is served by that server with the smallest index. For a fixed realization of the queueing system, think of assigning the customers when they arrive to the server who will eventually serve

them. Let $\underset{\sim}{W}_n = (W_{n1},\ldots,W_{ns})$ be the vector-valued process whose components W_{ni} denote the workload of the server with the ith lightest load just prior to the arrival of the nth customer. Thus W_{n1} is the actual waiting time of the nth customer. The $\underset{\sim}{W}_n$ sequence can be generated recursively as follows:

$$\underset{\sim}{W}_0 = \underset{\sim}{0}$$

$$\underset{\sim}{W}_{n+1} = [F(\underset{\sim}{W}_n + \underset{\sim}{V}_{n-1} - \underset{\sim}{U}_n)]^+,$$

where $\underset{\sim}{V}_{n-1} = (v_{n-1}, 0,\ldots,0)$, $\underset{\sim}{U}_n = (u_n,\ldots,u_n)$, $[\underset{\sim}{x}]^+ = (x_1^+,\ldots,x_s^+)$ for $\underset{\sim}{x} = (x_1,\ldots,x_s) \in \mathbb{R}^s$, and $F:\mathbb{R}^s \to \mathbb{R}^s$ rearranges the components of its argument in ascending order. It is clear from this representation of $\underset{\sim}{W}_n$ that $\{\underset{\sim}{W}_n : n \geq 0\}$ is a Markov process with state space $E = \{\underset{\sim}{x} \in \mathbb{R}^s : 0 \leq x_1 \leq \cdots \leq x_s\}$. The condition of Whitt mentioned above guarantees that the

$$P\{\underset{\sim}{W}_n = \underset{\sim}{0} \text{ infinitely often}\} = 1$$

and that the expected time between visits to $\underset{\sim}{0}$ is finite. This condition allows us to define the sequences of r.v.'s $\{\alpha_k : k \geq 1\}$, $\{\beta_k : k \geq 0\}$, and $\{\underset{\sim}{V}_k : k \geq 1\}$ just as in the case of the single-server queue. In this case customers numbered β_k arrive to find <u>all</u> servers idle.

Thus we can again proceed to decompose the queueing system into independent b.c.'s just as was done in the case $s = 1$. Now a b.c. is defined as an interval of time during which at least one

server is busy. In fact we can take over the same notation. However, in this case we have no closed expression for $E\{\alpha_k\}$ and thus we have no alternative but to carry out a simulation. Again $E\{\eta_k\} = \mu^{-1} E\{\alpha_k\}$, $E\{\xi_k\} = \lambda^{-1} E\{\alpha_k\}$, $E\{\nu_k\} = \lambda^{-1}(1-\rho) E\{\alpha_k\}$, and $E\{Y_k^{(1)}\} = E\{W\} E\{\alpha_k\}$ where W is again the stationary waiting time and $Y_k^{(1)}$ is the sum of the waiting times in the kth b.p. There is no closed expression for $E\{W\}$ and again simulation seems like the only recourse.

5. A TWO-STEP PROCEDURE FOR ESTIMATING $E\{f(W)\}$

Let f be some positive measurable function on the customer waiting time and let

$$Y_k = \sum_{j=0}^{\alpha_k - 1} f(W_{\beta_{k-1}+j}).$$

We know from the general theory that $\{Y_k : k \geq 1\}$ is a sequence of i.i.d. r.v.'s and that $E\{Y_1\} = E\{f(W)\} E\{\alpha_1\}$. This last relation is at the heart of the method to be proposed. By observing the sequences $\{Y_k : k \geq 1\}$ and $\{\alpha_k : k \geq 1\}$ and obtaining a $100(1-\gamma_1)\%$ confidence interval for $E\{Y_1\}$ and a $100(1-\gamma_2)\%$ confidence interval for $E\{\alpha_1\}$, it is possible to obtain a $100(1-\gamma_1-\gamma_2)\%$ confidence interval for $E\{f(W)\}$ using (4). Two questions naturally arise: given a desired $100(1-\gamma)\%$ confidence interval for $E\{f(W)\}$, how should γ_1 and γ_2 be chosen, and what effect do these choices have on the required sample size N ? It is these questions which we address in this section. Given a desired γ, we develop a two-step procedure in which the simulator chooses γ_1, γ_2, and N after obtaining preliminary estimates for system parameters in an initial "short" simulation run.

The full procedure consists of two independent simulation runs, the first for a total of m busy cycles and the second for N busy cycles, where N is to be determined. The initial run is a short run which serves an a planning guide for the much longer second run. The first run provides the simulator with estimates of confidence interval lengths for $E\{f(W)\}$ which could be obtained in the second run for various choices of γ_1, γ_2, and N. For each N, the estimated length is minimized over all possible values of γ_1 and γ_2 with $\gamma_1 + \gamma_2 = \gamma$, and the minimal length is expressed as a function of N. The simulator is thus able to consider tradeoffs between the length of confidence interval, level of confidence, and cost of computer resources.

Let $(Y_1,\ldots,Y_m)$ and $(\alpha_1,\ldots,\alpha_m)$ be observations during the initial run, and define

$$\bar{Y} = \frac{1}{m} \sum_{j=1}^{m} Y_j ,$$

$$\bar{\alpha} = \frac{1}{m} \sum_{j=1}^{m} \alpha_j ,$$

$$s_Y = \left[\frac{1}{m-1} \sum_{j=1}^{m} (Y_j-\bar{Y})^2\right]^{1/2},$$

$$s_\alpha = \left[\frac{1}{m-1} \sum_{j=1}^{m} (\alpha_j-\bar{\alpha})^2\right]^{1/2}.$$

Given these statistics, the simulator can estimate the size of confidence intervals for $E\{Y_1\}$ and $E\{\alpha_1\}$ which could be obtained in a second run, for various values of N, γ_1, and γ_2. If the second run were to be made with sample size N, $\gamma_1 = x$, and $\gamma_2 = \gamma - x$,

then estimates for $100(1-\gamma_1)\%$ and $100(1-\gamma_2)\%$ confidence intervals for $E\{Y_1\}$ and $E\{\alpha_1\}$ are

$$[\bar{Y} - z_{1-x/2}\, s_Y/\sqrt{N},\ \bar{Y} + z_{1-x/2}\, s_Y/\sqrt{N}$$

and

$$[\bar{\alpha} - z_{1-\gamma/2 + x/2}\, s_\alpha/\sqrt{N},\ \bar{\alpha} + z_{1-\gamma/2 + x/2}\, s_\alpha/\sqrt{N}].$$

Using (4), an estimate for the length of a $100(1-\gamma)\%$ confidence interval for $E\{f(W)\}$ is therefore

$$\ell(N,\gamma,x) = \frac{\bar{Y} + z_{1-x/2} s_Y/\sqrt{N}}{[\bar{\alpha} - z_{1-\gamma/2+x/2} s_\alpha/\sqrt{N}]^+} - \frac{[\bar{Y} - z_{1-x/2} s_Y/\sqrt{N}]^+}{\bar{\alpha} + z_{1-\gamma/2+x/2} s_\alpha/\sqrt{N}} .$$

For a fixed N and γ, it is desirable to choose that x which minimizes $\ell(N,\gamma,x)$. Using the expansion

$$\frac{1}{a+b} = \frac{1}{a} \sum_{j=0}^{\infty} (-1)^j \left(\frac{b}{a}\right)^j ,$$

we have, for large N, the approximation

$$\ell(N,\gamma,x) \approx F(N,\gamma,x) = \frac{2 s_Y}{\bar{\alpha}\sqrt{N}}\, z_{1-x/2} + \frac{2\bar{Y}\, s_\alpha}{\bar{\alpha}^2\sqrt{N}}\, z_{1-\gamma/2 + x/2}. \qquad (6)$$

We thus consider the problem of choosing x so as to minimize $F(N,\gamma,x)$, subject to $0 < x < \gamma$.

Recalling that $z_y = \Phi^{-1}(y)$ for $0 < y < 1$, it may be easily shown that $F(N,\gamma,x)$ is a strictly convex continuous function of x for $0 < x < \gamma$ and that $F(N,\gamma,x) \to \infty$ as $x \to 0$ or $x \to \gamma$. It follows that a unique minimum is obtained at $x = x_0$, the unique root in $(0,\gamma)$ of the equation $\frac{\partial}{\partial x} F(N,\gamma,x) = 0$.

Now

$$\frac{\partial}{\partial x} F(N,\gamma,x) = -\frac{s_\gamma}{\bar{\alpha}\sqrt{N}} \frac{1}{\phi(\Phi^{-1}(1-x/2))} + \frac{\bar{Y}s_\alpha}{\bar{\alpha}^2\sqrt{N}} \frac{1}{\phi(\Phi^{-1}(1-\gamma/2+x/2))}$$

$$= -\frac{\sqrt{2\pi}s_\gamma}{\bar{\alpha}\sqrt{N}} \exp\{\tfrac{1}{2} z^2_{1-x/2}\} + \frac{\sqrt{2\pi}\bar{Y}s_\alpha}{\bar{\alpha}^2\sqrt{N}} \exp\{\tfrac{1}{2} z^2_{1-\gamma/2+x/2}\}.$$

Setting this equal to zero, we obtain

$$z^2_{1-\gamma/2+x_0/2} - z^2_{1-x_0/2} = 2\ell n \frac{\bar{\alpha}s_\gamma}{\bar{Y}s_\alpha} \equiv c. \tag{7}$$

Note that the solution x_0 obtained in (7) is independent of N. That is, for a fixed γ, there is a unique choice $\gamma_1 = x_0$ which is optimal for every value of N. The estimate of the length of the confidence interval for $E\{f(W)\}$ is obtained from (6):

$$\ell_0(N,\gamma) = \ell(N,\gamma,x_0) \doteq D(\gamma)/\sqrt{N} \tag{8}$$

where

$$D(\gamma) = \frac{2s_\gamma}{\bar{\alpha}} z_{1-x_0/2} + \frac{2\bar{Y}s_\alpha}{\bar{\alpha}_2} z_{1-\gamma/2+x_0/2}. \tag{9}$$

Given (8) and (9), the simulator can make a rational choice of N and γ based on a consideration of computer time available, desired level of confidence, and desired length of confidence interval.

Table 1 shows, for several values of γ, the quantity $z^2_{1-\gamma/2 + x/2} - z^2_{1-x/2}$ as a function of γ and $h \equiv x/\gamma$. For a given value of c, computed after the initial simulation run, and of γ, Table 1 may thus be used to find that value x_0 solving (7). $D(\gamma)$ can then be computed from (9) using standard tables for the normal distribution. These computations are illustrated in Section 6.

It should be noted that (8) and (9) give only an estimate of the length of the confidence interval which could be obtained in the second simulation experiment of N busy cycles, this estimate being made at the end of the first experiment. The actual confidence interval is computed at the end of the second experiment, using confidence intervals for $E\{Y_1\}$ and $E\{\alpha_1\}$ together with (4).

It is interesting to contrast the method outlined above with classical sequential estimation procedures; cf. ANSCOMBE (1953). In the classical procedures, one generally desires to compute a confidence interval of a fixed length for a single unknown parameter. The interval length and confidence level are fixed at the start of the procedure, and the number of observations required is then dictated by the experimental observations and a sequential stopping rule, beyond the control of the experimentor. In contrast, the two-step procedure outlined in this section deals with estimation of the ratio of two unknown parameters, so that one must consider the problem of minimizing the interval length for the ratio over all possible lengths for the individual parameters. Furthermore, any

TABLE 1

VALUES OF $z^2_{1-\gamma/2 + x/2} - z^2_{1-x/2}$ AS A FUNCTION OF γ AND $h \equiv \frac{x}{\gamma}$

h \ γ	.005	.01	.02	.025	.05	.1	.2
.05	-5.4	-5.4	-5.3	-5.3	-5.2	-5.1	-4.9
.10	-4.0	-4.0	-3.9	-3.9	-3.9	-3.8	-3.6
.15	-3.2	-3.1	-3.1	-3.1	-3.0	-2.9	-2.8
.20	-2.5	-2.5	-2.5	-2.5	-2.5	-2.4	-2.2
.25	-2.0	-2.0	-2.0	-1.9	-1.9	-1.8	-1.8
.30	-1.5	-1.5	-1.5	-1.5	-1.5	-1.4	-1.3
.35	-1.2	-1.1	-1.1	-1.1	-1.1	-1.0	-1.0
.40	-0.7	-0.7	-0.7	-0.7	-0.7	-0.7	-0.6
.45	-0.4	-0.4	-0.4	-0.4	-0.4	-0.4	-0.3
.50	0.0	0.0	0.0	0.0	0.0	0.0	0.0
.55	0.4	0.4	0.4	0.4	0.4	0.4	0.3
.60	0.7	0.7	0.7	0.7	0.7	0.7	0.6
.65	1.2	1.1	1.1	1.1	1.1	1.0	1.0
.70	1.5	1.5	1.5	1.5	1.5	1.4	1.3
.75	2.0	2.0	2.0	1.9	1.9	1.8	1.8
.80	2.5	2.5	2.5	2.5	2.5	2.4	2.2
.85	3.2	3.1	3.1	3.1	3.0	2.9	2.8
.90	4.0	4.0	3.9	3.9	3.9	3.8	3.6
.95	5.4	5.4	5.3	5.3	5.2	5.1	4.9

decisions involving the desired degree of confidence, the interval length, and the ultimate number of observations are deferred until the end of the first-step experiment, when preliminary estimates for the parameters are available. The first step thus provides a basis for the economic decisions which must be made by the simulator in using limited computer resources.

6. SOME NUMERICAL ILLUSTRATIONS FOR THE M/M/1 QUEUE

In this section, we give numerical examples which combine the statistical techniques of Sections 2 and 5 with the queueing results of Sections 3 and 4. The M/M/1 queue is used for illustration because its theoretical properties are well-known and provide a means of comparison with the simulation results. In this case, the inter-arrival and service times are distributed as exponential r.v.'s with parameters λ and μ respectively.

From Section 3, we know that the sequences $\{\alpha_k : k \geq 1\}$, $\{\eta_k : k \geq 1\}$, $\{\xi_k : k \geq 1\}$, $\{\nu_k : k \geq 1\}$, $\{Y_k^{(1)} : k \geq 1\}$, $\{Y_k^{(2)} : k \geq 1\}$, and $\{Y_k^{(3)} : k \geq 1\}$ each consist of i.i.d. r.v.'s and are thus amenable to statistical analysis. In order to demonstrate the power and versatility of our methods, we define five additional sequences of i.i.d. r.v.'s. For $k \geq 1$ let

$$Y_k^{(4)} = \sum_{j=0}^{\alpha_k - 1} f_4(W_{\beta_{k-1}+j}) \equiv \sum_{j=0}^{\alpha_k - 1} (W_{\beta_{k-1}+j})^2,$$

$$Y_k^{(5)} = \sum_{j=0}^{\alpha_k - 1} f_5(W_{\beta_{k-1}+j}) \equiv \sum_{j=0}^{\alpha_k - 1} [(W_{\beta_{k-1}+j} - .1)^+]^{1/2},$$

$$Y_k^{(6)} = \sum_{j=0}^{\alpha_k - 1} f_6(W_{\beta_{k-1}+j}) \equiv \sum_{j=0}^{\alpha_k - 1} I_{\{W_{\beta_{k-1}+j} \geq .2\}},$$

$$Y_k^{(7)} = f_7(\alpha_k) \equiv \begin{cases} 0, & \alpha_k = 1,2 \\ 5, & \alpha_k = 3,4,5 \\ 10, & \text{otherwise,} \end{cases}$$

$$Y_k^{(8)} = f_8(\alpha_k) \equiv \begin{cases} 1, & \alpha_k = 1,2 \\ 0, & \text{otherwise,} \end{cases}$$

where the indicator function I_Λ takes on the value 1 on the set Λ and 0 otherwise. The function f_4 is useful for estimating the second moment and variance of the stationary waiting time W, since $E\{f_4(W)\} = E\{W^2\}$. The function f_5 can be interpreted as a penalty function on the waiting time of the customer. The function f_6 provides a means for estimating the tail of the stationary waiting time distribution, since $E\{f_6(W)\} = P\{W \geq .2\}$. The function f_7 can be interpreted as a penalty or cost function on the number of customers served during a busy period. This might be appropriate, for example, if servers are to be relieved at the end of each busy period and a penalty must be paid when a server is required to serve an excess number of customers without relief. Finally, the function f_8 provides information on the distribution of α_1, since $E\{f_8(\alpha_1)\} = P\{\alpha_1 = 1 \text{ or } \alpha_1 = 2\}$.

In order to illustrate the computation of confidence intervals for the various parameters of interest, a single-step simulation run was implemented with $\lambda = 5$, $\mu = 10$, and $N = 2000$ busy cycles (demonstration of the two-step procedure of Section 5 follows later). That is, 2000 observations were made from each of the sequences given above.

Now we have seen earlier that if $\{X_1,\ldots,X_N\}$ is a sample of i.i.d. r.v.'s with sample mean $\bar{X}$, sample standard deviation s, and finite second moments, then $[\bar{X} - z_{1-\gamma/2}\, s/\sqrt{N},\ \bar{X} + z_{1-\gamma/2}\, s/\sqrt{N}]$ is an approximate $100(1-\gamma)\%$ confidence interval for $E\{X_1\}$ for large N. In this manner, we obtain approximate confidence intervals, shown in Table 2, for $E\{\alpha_1\}$, $E\{\eta_1\}$, $E\{\xi_1\}$, $E\{v_1\}$, and $E\{Y_1^{(n)}\}$, $n = 1,2,\ldots,8$, based on the observed samples of 2000 (The moment condition is satisfied because all of the moments of u_1 and v_1 are finite.) Given approximate confidence intervals for $E\{Y_1^{(1)}\}$, $E\{Y_1^{(4)}\}$, $E\{Y_1^{(5)}\}$, $E\{Y_1^{(6)}\}$, and $E\{\alpha_1\}$, we use (4) to obtain approximate confidence intervals for $E\{W\}$, $E\{f_4(W)\}$, $E\{f_5(W)\}$, and $E\{f_6(W)\}$, since

$$E\{W\} = \frac{E\{Y_1^{(1)}\}}{E\{\alpha_1\}} ,$$

$$E\{f_4(W)\} = \frac{E\{Y_1^{(4)}\}}{E\{\alpha_1\}} ,$$

$$E\{f_5(W)\} = \frac{E\{Y_1^{(5)}\}}{E\{\alpha_1\}} ,$$

and

$$E\{f_6(W)\} = \frac{E\{Y_1^{(6)}\}}{E\{\alpha_1\}} .$$

Given an approximate confidence interval for $E\{W\}$, we use (2) to obtain an approximate confidence interval for $E\{Q\} = \lambda E\{W\} + \rho$. Finally, given approximate confidence intervals for $E\{W\}$ and for $E\{f_4(W)\} = E\{W^2\}$, we use (1) and (3) to obtain an approximate confidence interval for the standard deviation $\sigma\{W\} = [E\{W^2\} - (E\{W\})^2]^{1/2}$.

Table 2 summarizes the results of the simulation run. For each entry through $E\{Y_1^{(8)}\}$, the value given in the table for the "point estimate" is the value of the sample mean. For each of the remaining entries, the point estimate is the appropriate function of the other point estimates; e.g., the estimate for $E\{W\}$ is the ratio of the estimates for $E\{Y_1^{(1)}\}$ and $E\{\alpha_1\}$.

Table 3 shows the point estimates and 90% confidence intervals for $E\{W\}$ in ten replications of the experiment. As can be seen, for these runs, each of the ten intervals surrounds the true mean $E\{W\} = .1$.

We next demonstrate the two-step procedure of Section 5 for obtaining a confidence interval for $E\{f(W)\}$. For illustrative purposes, we let f be the identity function, so that we estimate $E\{W\}$.

In order to illustrate the procedure a first-step run of 500 busy cycles was made. Estimates obtained from this run are

$$\bar{Y} = .178 ,$$

$$\bar{\alpha} = 1.95 ,$$

$$s_Y = .749 ,$$

$$s_\alpha = 2.20 ,$$

so that

$$c = 2 \ln \left(\frac{\bar{\alpha} s_Y}{\bar{Y} s_\alpha}\right) = 2.63.$$

The optimal value $\gamma_1 = x_0$ may now be obtained from Table 1. For several values of γ, Table 4 shows the optimal γ_1, $\gamma_2 = \gamma - \gamma_1$, and the associated $D(\gamma)$. For a second-step simulation of N busy cycles, $\ell_0(N,\gamma) = D(\gamma)/\sqrt{N}$ is the estimate for the minimum length

TABLE 2

SIMULATION RESULTS FOR THE M/M/1 QUEUE

(arrival rate $\lambda = 5$; service rate $\mu = 10$; N = 2000 observed busy cycles)

Parameter	Theoretical Value	Point Estimate	Confidence Interval	Level of Confidence
$E\{\alpha_1\}$	2.000	2.110	[1.994, 2.226]	95%
$E\{\eta_1\}$	0.200	0.215	[.199, .231]	95%
$E\{\xi_1\}$	0.400	0.416	[.396, .435]	95%
$E\{\nu_1\}$	0.200	0.201	[.192, .210]	95%
$E\{Y_1^{(1)}\} = E\{\sum_{j=0}^{\alpha_1-1} W_j\}$	0.200	0.232	[.187, .277]	95%
$E\{Y_1^{(2)}\} = E\{\int_0^{\xi_1} Q(s)ds\}$	0.400	0.447	[.387, .507]	95%
$E\{Y_1^{(3)}\} = E\{\int_0^{\xi_1} W(s)ds\}$	0.040	0.045	[.038, .053]	95%
$E\{Y_1^{(4)}\} = E\{\sum_{j=0}^{\alpha_1-1} (W_j)^2\}$	0.080	0.096	[.066, .127]	95%
$E\{Y_1^{(5)}\} = E\{\sum_{j=0}^{\alpha_1-1} \sqrt{(W_j-.1)^+}\}$	0.240	0.280	[.226, .333]	95%
$E\{Y_1^{(6)}\} = E\{\sum_{j=0}^{\alpha_1-1} I_{\{W_j \geq .2\}}\}$	0.368	0.438	[.357, .519]	95%
$E\{Y_1^{(7)}\} = E\{f_4(\alpha_1)\}$	1.225	1.375	[1.248, 1.502]	95%
$E\{Y_1^{(8)}\} = P\{\alpha_1=1 \text{ or } \alpha_1=2\}$	0.815	0.792	[.774, .810]	95%
$E\{W\}$	0.100	0.110	[.084, .139]	90%
$E\{f_4(W)\} = E\{W^2\}$	0.040	0.046	[.030, .064]	90%
$E\{f_5(W)\} = E\{\sqrt{(W-.1)^+}\}$	0.120	0.133	[.102, .167]	90%
$E\{f_6(W)\} = P\{W \geq .2\}$	0.184	0.208	[.160, .260]	90%
$E\{Q\}$	1.000	1.050	[.920, 1.195]	90%
$\sigma\{W\}$	0.173	0.182	[.101, .238]	80%

TABLE 3

ESTIMATES FOR E(W) IN TEN SIMULATION REPLICATIONS

($\lambda = 5$; $\mu = 10$; N = 2000 observed busy cycles; level of confidence $\geq$ 90%)

Replication	Point Estimate	Confidence Interval
1	0.110	[.084, .139]
2	0.091	[.071, .114]
3	0.095	[.075, .117]
4	0.111	[.075, .151]
5	0.096	[.073, .122]
6	0.100	[.077, .126]
7	0.092	[.071, .116]
8	0.099	[.074, .128]
9	0.096	[.073, .122]
10	0.090	[.068, .115]

TABLE 4

VALUES FOR γ_1, γ_2, AND $D(\gamma)$ AS A FUNCTION OF γ

(Based on a simulation run with $\lambda = 5$, $\mu = 10$, $m = 500$ observed busy cycles.)

γ	γ_1	γ_2	$D(\gamma)$
.2	.167	.033	1.50
.1	.082	.018	1.83
.05	.041	.009	2.11
.02	.0164	.0036	2.44
.01	.0081	.0019	2.67

TABLE 5

SIMULATION RESULTS FOR THE TWO-STEP PROCEDURE

($\lambda = 5$, $\mu = 10$, $N = 18{,}000$ observed busy cycles)

Parameter	Theoretical Value	Point Estimate	Confidence Interval	Level of Confidence
$E\{\alpha_1\}$	2.000	1.983	[1.932, 2.035]	99.64%
$E\{Y_1^{(1)}\}$	0.200	0.192	[0.175, 0.209]	98.36%
$E\{W\}$	0.100	0.097	[0.086, 0.108]	98.00%

confidence interval for $E(W)$, and Figure 1 shows $\ell_0(N,\gamma)$ as a function of N and γ. For example if N is chosen to be 4000 (about a 5 second run on the IBM 360/67) and $\gamma = .05$, then the estimated confidence interval length is .033, for a deviation of ±18% about the point estimate for $E\{W\}$.

Using Figure 1 as a guide, it was decided that the second experiment be run with $N = 18{,}000$ and $\gamma = .02$ (98% confidence). Table 5 shows the actual results of this run. The confidence interval obtained for $E(W)$ is of length 0.022, compared to a length of 0.018 estimated at the end of the first step.

Suppose now that one desires to make comparative inferences about two different queueing systems. For example, one might be interested in comparing the mean stationary waiting times $E\{W^{(1)}\}$ and $E\{W^{(2)}\}$ for systems with $\mu = 10$ and $\mu = 16$, each having an arrival rate $\lambda = 5$. One method of comparison is to obtain a confidence interval on the difference in the mean waiting times. This can be done by computing confidence intervals on the means $E\{W^{(1)}\}$ and $E\{W^{(2)}\}$ separately, then using (1) for the difference.

To illustrate, a run of 18,000 busy cycles was made with $\mu = 16$, resulting in a 98% confidence interval for $E\{W^{(2)}\}$ of [.025, .030] (γ_1 and γ_2 were chosen the same as earlier for $\mu = 10$). Combining this with the previously computed interval for $E\{W^{(1)}\}$, we obtain the interval [.056, .083] for $E\{W^{(1)}\} - E\{W^{(2)}\}$. That is, with confidence not less than 96%, the mean stationary waiting time for $\mu = 10$ exceeds that for $\mu = 16$ by at least .056 but not more than .083. Statements of this type would be useful for assessing the relative value of different proposed system modifications.

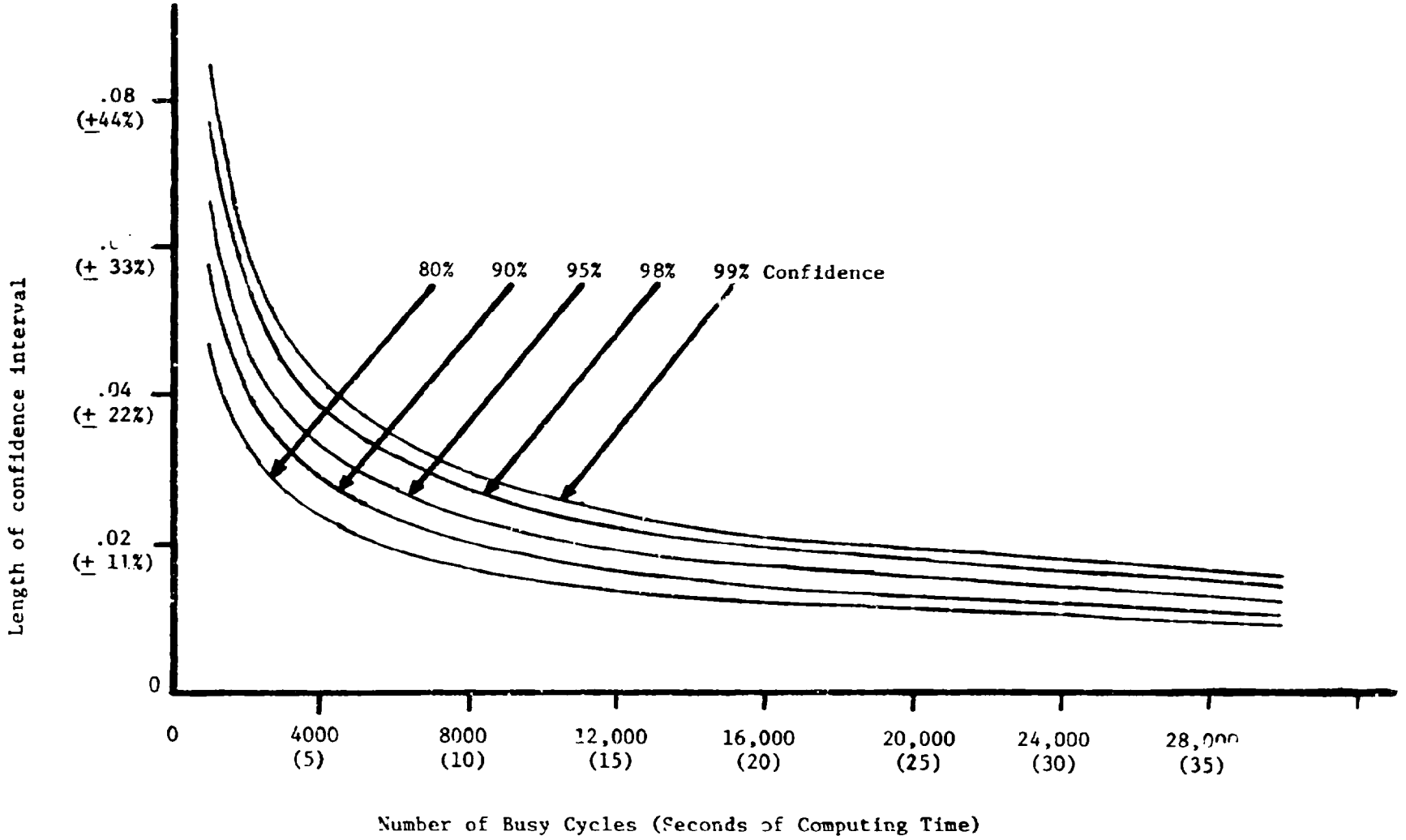


Figure 1. ESTIMATED LENGTH OF CONFIDENCE INTERVAL FOR THE MEAN WAITING TIME AS A FUNCTION OF NUMBER OF BUSY CYCLES AND LEVEL OF CONFIDENCE (based on a first-step run of 500 busy cycles with arrival rate $\lambda = 5$ and service rate $\mu = 10$)

Suppose one wishes to study the sensitivity of the queueing system over a range of parameter values, say for $3 \leq \lambda \leq 5$ with $\mu = 10$. We illustrate again for the mean waiting time $E\{W(\lambda)\}$. If one assumes that $E\{W(\lambda)\}$ is approximately a linear function of λ for $3 \leq \lambda \leq 5$, then (5) may be used to obtain an approximate confidence band about the function $E\{W(\lambda)\}$ over that interval. To illustrate, a run of 18,000 busy cycles was made for the case $\lambda = 3$ and $\mu = 10$, and a 98% confidence interval obtained for the mean waiting time is [.037, .045]. An interval [.086, .108] was previously obtained for $\lambda = 5$, and $\mu = 10$. Then under the linear assumption for $E\{W(\lambda)\}$ between $\lambda = 3$ and $\lambda = 5$, a 96% confidence band for $E\{W(\lambda)\}$ is shown in Figure 2 together with the true function $E\{W(\lambda)\}$. That is, with at least 96% confidence, $.037 + (.049)(\lambda-3)/2 \leq E\{W(\lambda)\} \leq .045 + (.063)(\lambda-3)/2$ for all $3 \leq \lambda \leq 5$.

The same method could be combined with the previous method for comparing alternative systems. For example, if $E\{W(\lambda,\mu)\}$ is the mean waiting time for a system with parameters λ and μ, and if μ_1 and μ_2 are two alternative values for μ, one might wish a confidence band for $g(\lambda) \equiv [E\{W(\lambda,\mu_1)\} - E\{W(\lambda,\mu_2)\}]$ over a range $\lambda_1 \leq \lambda \leq \lambda_2$. That is, inferences would be made on the sensitivity of the system differences over the given range for λ, based on the assumption of linearity over this range. This is done by making runs at only four different parameter settings: (λ_1, μ_1), (λ_1,μ_2), (λ_2,μ_1), and (λ_2,μ_2). Figure 3 illustrates our results.

This graph illustrates one particular experimental realization, based on observations at the two endpoints $\lambda = 3$ and $\lambda = 5$. In at least 98% of such realizations, the confidence limits at either endpoint will surround the true value. In at least 96% of the realizations, the confidence limits at all points $3 \leq \lambda \leq 5$ will surround the true values. Thus, for example, a 98% confidence interval at $\lambda = 4$ is [.062, .076] in this realization.

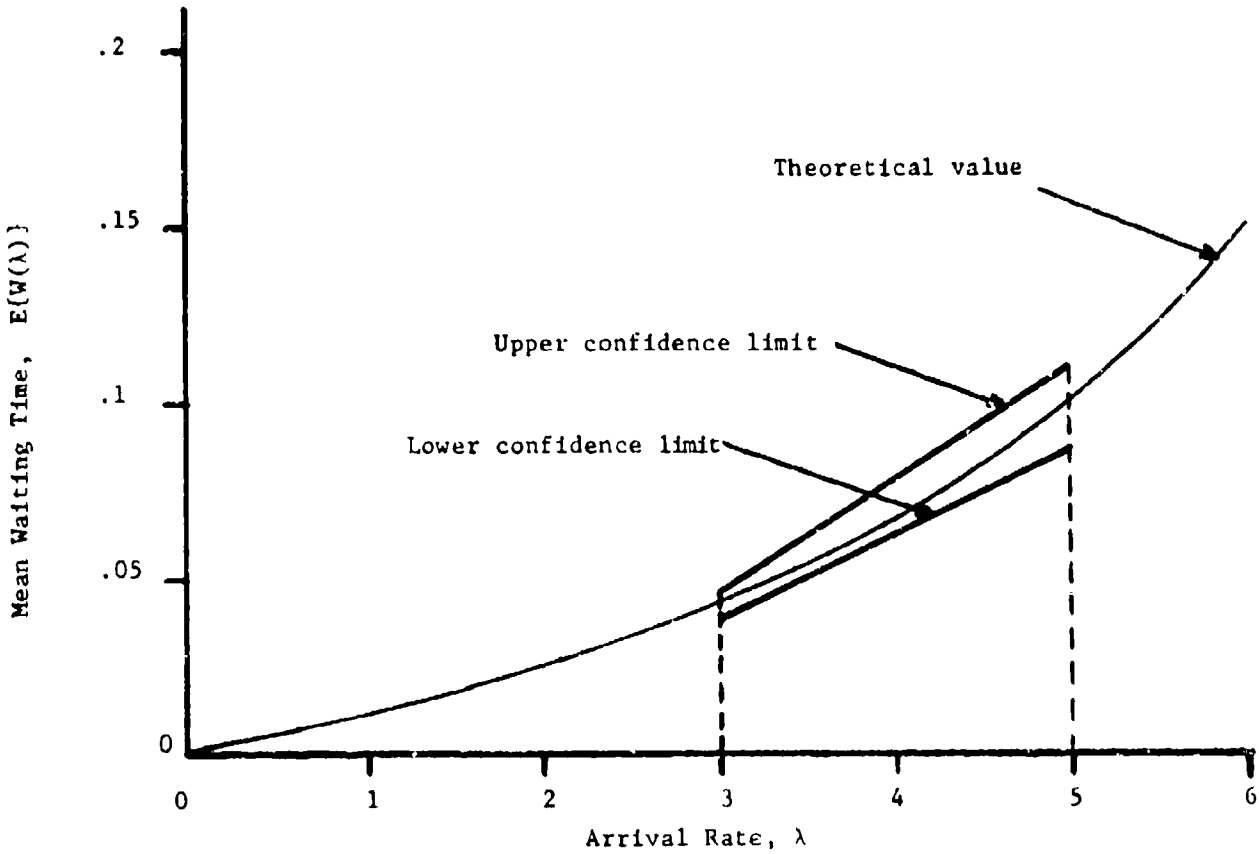


Figure 2. A 96% CONFIDENCE BAND FOR THE MEAN WAITING TIME, AS A FUNCTION OF THE ARRIVAL RATE λ OVER $3 \leq \lambda \leq 5$. (Service rate $\mu = 10$; N = 18,000 observed busy cycles.)

This graph illustrates one particular experimental realization, based on observations at the two endpoints $\lambda = 3$ and $\lambda = 5$. In at least 96% of such realizations, the confidence limits at either endpoint will surround the true value. In at least 92% of the realizations, the confidence limits at all points $3 \leq \lambda \leq 5$ will surround the true values. Thus, for example, a 92% confidence interval at $\lambda = 4$ is [.039, .057] in this realization.

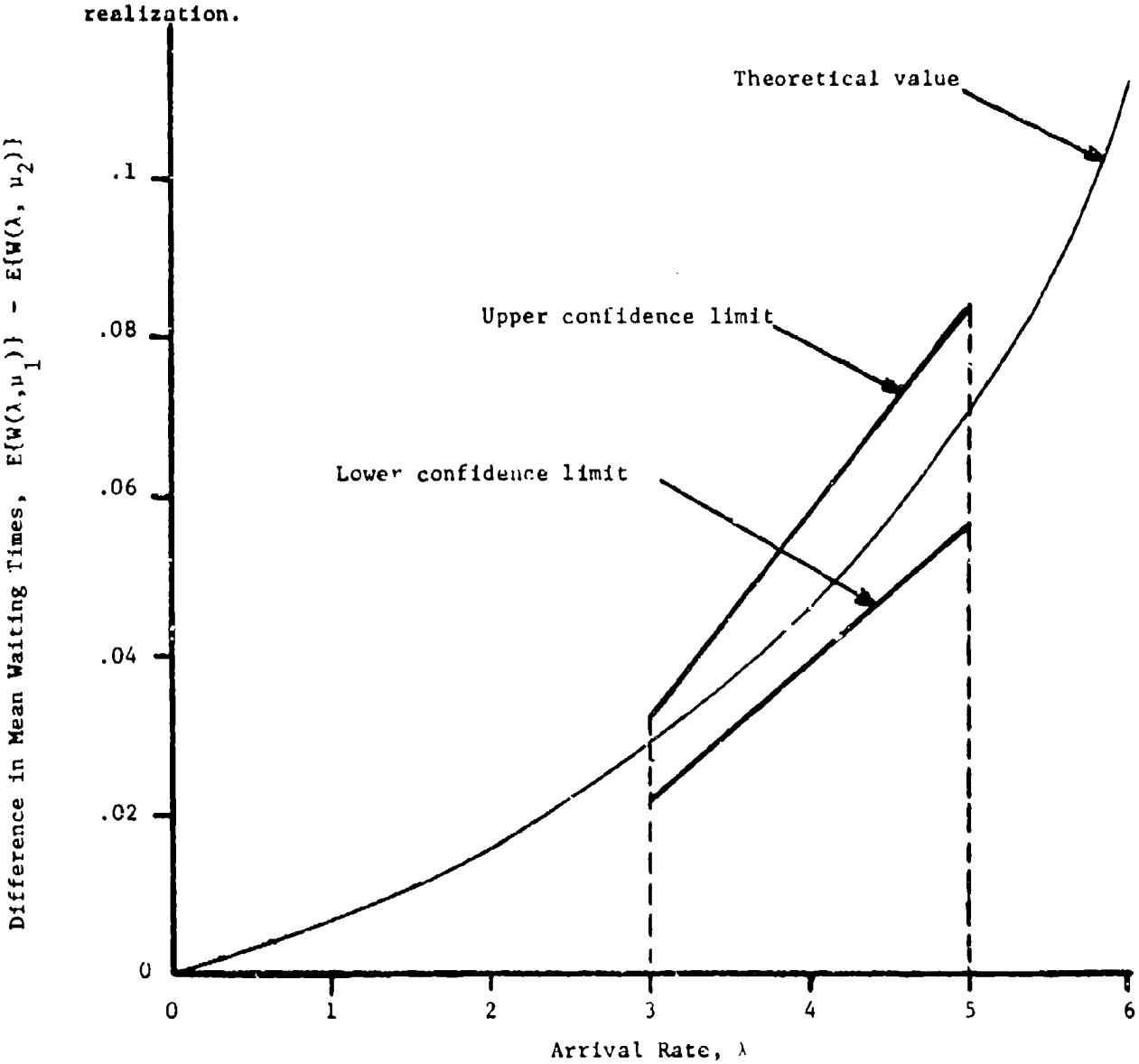


Figure 3. A 92% CONFIDENCE BAND FOR THE DIFFERENCE IN MEAN WAITING TIMES BETWEEN TWO SYSTEMS WITH DIFFERENT SERVICE RATES, AS A FUNCTION OF THE ARRIVAL RATE λ OVER $3 \leq \lambda \leq 5$. (Service rates $\mu_1 = 10$, $\mu_2 = 16$; N = 18,000 observed busy cycles.)

The key assumption is of course the linear relationship over the range of interest. This assumption is fairly accurate in the examples given, as seen in Figures 2 and 3. In future work, we intend to explore the possibility of obtaining confidence bands which are non-linear. It is believed that the theory of nonlinear regression may be brought to bear on this problem.

7. COMPARISON WITH AN ALTERNATIVE METHOD

In previous sections, we have shown how to obtain a confidence interval for $E(W)$ roughly centered about

$$\bar{W}_N \equiv \frac{\frac{1}{N}\sum_{k=1}^{N} Y_k^{(1)}}{\frac{1}{N}\sum_{k=1}^{N} \alpha_k} = \frac{1}{\beta_N}\sum_{j=0}^{\beta_N - 1} W_j \quad ,$$

where N is the number of busy cycles observed. We know by the strong law of large numbers that as $N \to \infty$,

$$\frac{1}{N}\sum_{k=1}^{N} Y_k^{(1)} \to E\{Y_1^{(1)}\} = E\{\alpha_1\}\ E\{W\} \quad \text{a.e.}$$

and

$$\frac{1}{N}\sum_{k=1}^{N} \alpha_k \to E\{\alpha_1\} \quad \text{a.e.}$$

so that

$$\bar{W}_N \to E\{W\} \quad \text{a.e.} \ .$$

Hence, the estimator $\bar{W}_N$ is consistent. Given any $\epsilon > 0$, the

$$\lim_{N\to\infty} P\{|\bar{W}_N - E\{W\}| \leq \epsilon\} = 1.$$

An alternative estimator for $E\{W\}$ is

$$\bar{W}'_M = \frac{1}{M} \sum_{j=0}^{M-1} W_j.$$

For many queueing systems, e.g. the M/G/1 and GI/M/1 queues, $\bar{W}'_M$ is that estimator among the class of estimators

$$\bar{W}'_{M,m} = \frac{1}{M-m} \sum_{j=m}^{M-1} W_j \qquad m = 0,1,2,\ldots,M-1$$

which minimizes the "mean square error"

$$E\{(\bar{W}'_{M,m} - E\{W\})^2\}$$

for all large values of M, see BLOMQVIST (1970).

Now let N_0 be a fixed positive integer, and let

$$M_0 = E\{\alpha_1\} N_0.$$

The two estimators $\bar{W}_{N_0}$ and $\bar{W}'_{[M_0]}$ require, on the average, the same length of simulation, and it is of interest to compare their respective differences from $E\{W\}$. For a fixed $\epsilon_0 > 0$, define

$$p \equiv P\{|\bar{W}_{N_0} - E\{W\}| \leq \varepsilon_0\},$$

$$p' \equiv P\{|\bar{W}'_{[M_0]} - E\{W\}| \leq \varepsilon_0\}.$$

By replicating a number of simulation experiments, it is possible to obtain point estimates and confidence intervals for the parameters p and p'. To illustrate, 400 independent simulation replications were made for the M/M/1 example of Section 6, with $\lambda = 5$ and $\mu = 10$. In this case $E\{\alpha_1\} = 2$ and we set $N_0 = 100$ and $M_0 = 200$. Point estimates and confidence intervals for p and p' are shown in Table 6 for three different values of ε_0. We see from Table 6 that there are no significant differences between p and p', implying that one can generally expect comparable "accuracy" from the estimators $\bar{W}_{N_0}$ and $\bar{W}'_{[M_0]}$. Similar results are obtained for other values of N_0 and $M_0 = E\{\alpha_1\} N_0$. These findings lend respectability to the methods outlined in the previous sections, for those methods would be of questionable value if they were based on estimators less "accurate" than commonly used alternatives. Of course the main advantage of our methods over alternatives is the ability to use statistical analysis appropriate for sequences of i.i.d. r.v.'s.

TABLE 6

ESTIMATES FOR $p \equiv P\{|\bar{W}_{N_0} - E\{W\}| \leq \varepsilon_0$ AND $p' \equiv P\{|\bar{W}'_{[M_0]} - E\{W\}| \leq \varepsilon_0$

(400 replications with $\lambda = 5$, $\mu = 10$, $N_0 = 100$, and $M_0 = 200$; $E\{W\} = 0.1$)

ε_0	Point Estimates		90% Confidence Intervals	
	p	p'	p	p'
0.01	.220	.225	[.186, .254]	[.191, .259]
0.02	.455	.423	[.414, .496]	[.382, .464]
0.03	.625	.623	[.585, .665]	[.583, .663]

REFERENCES

[1] ANSCOMBE, F.J. (1953). Sequential estimation. J.R. Statist. Soc. B, 15, 1-21.

[2] BLOMQVIST, N. (1970). On the transient behavior of the GI/G/1 waiting times. Skand. Akt. tidskr. 118-129.

[3] CHUNG, K.L. (1968). A Course in Probability Theory, Harcourt, Brace and World, New York.

[4] IGLEHART, D.L. (1971). Functional limit theorems for the queue GI/G/1 in light traffic. Adv. Appl. Prob. 3, 269-281.

[5] KENNEDY, D.P. (1971). A note on the number of busy servers in a GI/G/s queue in light traffic. Research Report 96/DPK 5. Department of Prob. and Statist., The University, Sheffield.

[6] KIEFER, J. and WOLFOWITZ, J. (1955). On the theory of queues with many servers. Trans. Amer. Math. Soc. 78, 1-18.

[7] KIEFER, J. and WOLFOWITZ, J. (1956). On the characteristics of the general queueing process with applications to random walk. Ann. Math. Statist. 27, 147-161.

[8] LOYNES, R.M. (1962). The stability of a queue with non-independent interarrival and service times. Proc. Camb. Phil. Soc. 58, 497-520.

[9] WHITT, W. (1971). Embedded renewal processes in the GI/G/1 queue. Technical Report, Yale University. To appear.